Thot. New features to deal with larger corpora and longer sentences

OpenLab 2006

D .Ortiz, I. García-Varea, F. Casacuberta, L. Rodríguez and J. Tomás

March 30, 2006 – Trento, Italy

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

Index

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

What's Thot

- Current state of the art in SMT: phrase-based approach
- Current tasks: huge and difficult
- Available toolkit:
  - decoding(+training): Pharaoh (not GPL, not OpenSource)
- Thot
  - is a GPL toolkit to train PB Statistical Translation Models
  - is Open Source → customizable
  - first release delivered on August 2005: `http://www.info-ab.uclm.es/simd/software`
  - an improved version (planned by April 2006): `http://sourceforge.net`

Who can use it?

- Machine translation community:
  - MT researchers
  - university programs: PhD and graduated students
  - public institutions: European Union, regional governments
  - private companies
  - ...
- Linguistic community:
  - typically not familiar with mathematical details
  - non programming skills
  - very useful to obtain bilingual dictionaries automatically

Index

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

Phrase-based translation

- SMT is based on the source-channel model:

$$\arg\max_{e_1^I} Pr(e_1^I | f_1^J) = \arg\max_{e_1^I} Pr(e_1^I) \cdot Pr(f_1^J | e_1^I)$$

$$\begin{aligned} Pr(e_1^I) &: \text{ output language model} \\ Pr(f_1^J | e_1^I) &: \text{ inverse phrase-based translation model} \end{aligned}$$

- To model the relations at phrase-level (bilingual phrase alignments), a hidden variable $\tilde{\mathbf{a}} = \tilde{a}_1^K$ is introduced:

$$Pr(f_1^J | e_1^I) = \sum_{\tilde{\mathbf{a}}} Pr(\tilde{\mathbf{a}}, \tilde{f}_1^J | \tilde{e}_1^I) = \sum_{\tilde{\mathbf{a}}} Pr(\tilde{\mathbf{a}} | \tilde{e}_1^I) Pr(\tilde{f}_1^J | \tilde{\mathbf{a}}, \tilde{e}_1^I)$$

Modelling

- Typical assumptions to the models reduce them to phrase-based statistical dictionaries:

$$Pr(f_1^J|e_1^I) = \alpha(e_1^I) \sum_{\tilde{\mathbf{a}}} \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k})$$

- Maximum likelihood (ML) estimation: $\theta = \{p(\tilde{f}|\tilde{e})\}$:

$$\hat{\theta} = \arg\max_{\theta} p_\theta(f_1^J|e_1^I) = \arg\max_{\theta} \sum_{\tilde{\mathbf{a}}} \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k})$$

- Parameter estimation methods:
  - ML estimation via EM algorithm (correct)
  - $\Rightarrow$ From single-word alignment matrices (the "classical" approach, heuristic)

Index

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

Index

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

Toolkit description

Main tools (typically used in that order):

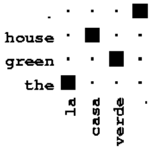
- Alignment combination ($\cup$,$\cap$,Sum,Symm.)$\rightarrow$ improvements in training
- Parameter estimation of a PB model from word alignment matrices:
  - Relative frequencies (RF): *phrase-extract*
  - *pseudo ML-estimation* (pML)
- Bilingual segmentation $\rightarrow$ post-processing step (e.g. to be used by finite state transducers)

Parameter estimation

- Bilingual phrases must be consistent with its corresponding word alignment matrix A as:([Och, 2002]):

$$\mathcal{BP}(f_1^J, e_1^I, A) = \\ \{(f_j^{j+m}, e_i^{i+n}) : \forall(i', j') \in A : j \leq j' \leq j+m \wedge i \leq i' \leq i+n\}$$

- Set of consistent bilingual phrases (right) given a word alignment matrix (left):



source phrase	target phrase
La	the
casa	house
verde	green
casa verde	green house
La casa verde	the green house
.	.
casa verde .	green house .
La casa verde .	the green house .

Parameter estimation: RF

- For every (f_1^J, e_1^I, A):
  1. Obtain the set $\mathcal{BP}(f_1^J, e_1^I, A)$
  2. $\forall(\tilde{f}, \tilde{e}) \in \mathcal{BP}(f_1^J, e_1^I, A)$ update counts:

$$count(\tilde{f}, \tilde{e}) = count(\tilde{f}, \tilde{e}) + 1$$

- Consequently $p(\tilde{f}|\tilde{e})$ is computed as:

$$p(\tilde{f}|\tilde{e}) = \frac{count(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} count(\tilde{f}, \tilde{e})}$$

Parameter estimation: RF

- For every (f_1^J, e_1^I, A):
  1. Obtain the set $\mathcal{BP}(f_1^J, e_1^I, A)$
  2. $\forall(\tilde{f}, \tilde{e}) \in \mathcal{BP}(f_1^J, e_1^I, A)$ update counts:

$$count(\tilde{f}, \tilde{e}) = count(\tilde{f}, \tilde{e}) + 1$$

- Consequently $p(\tilde{f}|\tilde{e})$ is computed as:

$$p(\tilde{f}|\tilde{e}) = \frac{count(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} count(\tilde{f}, \tilde{e})}$$

Disadvantage$\rightarrow$ bilingual phrases are not considered as part of complete bilingual segmentations

Parameter estimation: pML

- For every (f_1^J, e_1^I, A):
  1. Obtain the set $\mathcal{BP}(f_1^J, e_1^I, A)$
  2. Obtain the set $\mathcal{S}_{\mathcal{BP}(f_1^J, e_1^I, A)}$ of all partitions of the pair (f_1^J, e_1^I)
  3. $\forall(\tilde{f}, \tilde{e}) \in \mathcal{S}_{\mathcal{BP}(f_1^J, e_1^I, A)}$ update (fractional) counts:

$$count(\tilde{f}, \tilde{e}) += \frac{N(\tilde{f}, \tilde{e})}{|\mathcal{S}_{\mathcal{BP}(f_1^J, e_1^I, A)}|}$$

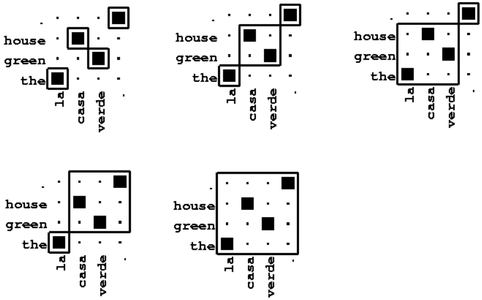
$N(\tilde{f}, \tilde{e})$: number of times that $(\tilde{f}, \tilde{e})$ occurs in $\mathcal{S}_{\mathcal{BP}(f_1^J, e_1^I, A)}$.

- Consequently $p(\tilde{f}|\tilde{e})$ is computed as:

$$p(\tilde{f}|\tilde{e}) = \frac{count(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}} count(\tilde{f}, \tilde{e})}$$

Parameter estimation: an example of pML estimation

Possible segmentations for a given word-alignment matrix:



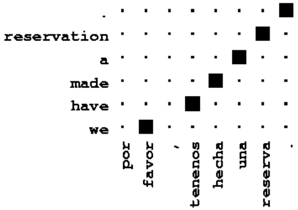
Parameter estimation: an example of pML estimation

Bilingual phrase counts for RF and pML estimation

$\tilde{f}$ — $\tilde{e}$	RF	pML	
La — the	1	3/5	⇐
casa — house	1	1/5	
verde — green	1	1/5	
casa verde — green house	1	1/5	
La casa verde — the green house	1	1/5	
. — .	1	3/5	⇐
casa verde . — green house .	1	1/5	
La casa verde . — the green house .	1	1/5	

Parameter estimation: RF vs. pML

Given the following word-aligned bilingual pair:



- RF estimation yields 22 phrase pairs
- pML estimation produces 31 segmentations using 21 phrase pairs
- The phrase pair *favor* → *we* is not part of any valid segmentation

Index

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

New release features

Limitations of current release of Thot:

- Most used corpora contain a huge amount of data $\rightarrow$ huge model $\rightarrow$ high memory requirements
- pML estimation $\rightarrow$ yields a high computational cost to obtain the set of bilingual segmentations
- Useful only for small task

New release features

The new version of Thot

- implements an efficient algorithm to compute the set of bilingual segmentations (for pML estimation)
- provides an incremental learning framework
- is not constrained by the size of the corpora
- provides an API for decoding with:
  - very efficient memory management: low cost data structures
  - an efficient retrieval of PB model probabilities: trie+caching

Index

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

Index

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

Experimental framework

- Experiments developed on the OpenLab 2006 task
- Training and test corpus pre-processed
- Default decoding and training parameters
- Training → comparison between

  Pharaoh : RF estimation
  Thot : pML estimation
- Decoding → Pharaoh decoder

Index

1 Introduction

2 Phrase-based translation (review)

3 Thot. A toolkit to train phrase-based models
- Toolkit description
- New release features

4 Experimental results
- Experiments
- Results

OpenLab corpus results

- Total training time:

  Pharaoh : ≈ 24 hours
  Thot : ≈ 7 hours

- Results on the OpenLab 2006 test-set

System	WER	CER	PER	NIST	BLEU
Pharaoh	42.97	33.96	34.18	8.89	0.41
Thot	41.62	33.55	33.27	8.94	0.42

Och, F. J. (2002).
Statistical Machine Translation: From Single-Word Models to Alignment Templates.
PhD thesis, Computer Science Department, RWTH Aachen, Germany.